Feature Screening via Distance Correlation Learning

Runze Li, Wei Zhong and Liping Zhu

The Pennsylvania State University, Xiamen University

& Shanghai University of Finance and Economics

March 4, 2013

Abstract

This paper is concerned with screening features in ultrahigh dimensional data analysis, which has become increasingly important in diverse scientific fields. We develop a sure independence screening procedure based on the distance correlation (DC-SIS, for short). The DC-SIS can be implemented as easily as the sure independence screening procedure based on the Pearson correlation (SIS, for short) proposed by Fan and Lv (2008). However, the DC-SIS can significantly improve the SIS. Fan and Lv (2008)

*Runze Li is Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111. Email: rli@stat.psu.edu. His research was supported by National Institute on Drug Abuse (NIDA) grant P50-DA10075 and National Natural Science Foundation of China (NNSFC) grant 11028103. Wei Zhong is the corresponding author and Assistant Professor of Wang Yanan Institute for Studies in Economics, Department of Statistics and Fujian Key Laboratory of Statistical Science, Xiamen University, China. Email: wxz123@psu.edu. His research was supported by a NIDA grant P50-DA10075 as a graduate research assistant during his graduate study, and by the NNSFC grant 71131008 (Key Project). Liping Zhu is Associate Professor of School of Statistics and Management, Shanghai University of Finance and Economics, China. Email: zhu.liping@mail.shufe.edu.cn. His research was supported by NNSFC grant 11071077 and a NIDA grant R21-DA024260. All authors equally contribute to this paper, and the authors are listed in the alphabetic order. The authors thank the Editor, the AE and reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIDA.

established the sure screening property for the SIS based on linear models, but the sure screening property is valid for the DC-SIS under more general settings including linear models. Furthermore, the implementation of the DC-SIS does not require model specification (e.g., linear model or generalized linear model) for responses or predictors. This is a very appealing property in ultrahigh dimensional data analysis. Moreover, the DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables. We establish the sure screening property for the DC-SIS, and conduct simulations to examine its finite sample performance. Numerical comparison indicates that the DC-SIS performs much better than the SIS in various models. We also illustrate the DC-SIS through a real data example.

Key words: Distance correlation, sure screening property, ultrahigh dimensionality, variable selection.

Running Head: Distance Correlation Based SIS

1. INTRODUCTION

Various regularization methods have been proposed for feature selection in high dimensional data analysis, which has become increasingly frequent and important in various research fields. These methods include, but are not limited to, the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001; Kim, Choi and Oh, 2008; Zou and Li, 2008), the LARS algorithm (Efron, Hastie, Johnstone and Tibshirani, 2004), the elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009), the adaptive LASSO (Zou, 2006) and the Dantzig selector (Candes and Tao, 2007). All these methods allow the number of predictors to be greater than the sample size, and perform quite well for high dimensional data.

With the advent of modern technology for data collection, researchers are able to collect ultrahigh dimensional data at relatively low cost in diverse fields of scientific research. The aforementioned regularization methods may not perform well for ultrahigh dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan, Samworth and Wu, 2009). These challenges call for new statistical modeling techniques for ultrahigh dimensional data. Fan and Lv (2008) proposed the SIS and showed that the Pearson correlation ranking procedure possesses a sure screening property for linear regressions with Gaussian predictors and responses. That is, all truly important predictors can be selected with probability approaching one as the sample size diverges to ∞. Hall and Miller (2009) extended Pearson correlation learning by considering polynomial transformations of predictors. To rank the importance of each predictor, they suggested a bootstrap procedure. Fan, Samworth and Wu (2009) and Fan and Song (2010) proposed a more general version of independent learning which ranks the maximum marginal likelihood estimators or the maximum marginal likelihood for generalized linear models. Fan, Feng and Song (2011) considered nonparametric independence screening in sparse ultrahigh dimensional additive models. They suggested estimating the nonparametric components marginally with spline approximation, and ranking the importance of predictors using the

magnitude of nonparametric components. They also demonstrated that this procedure possesses the sure screening property with vanishing false selection rate. Zhu, Li, Li and Zhu (2011) proposed a sure independent ranking and screening (SIRS) procedure to screen significant predictors in multi-index models. They further show that under linearity condition assumption on the predictor vector, the SIRS enjoys the ranking consistency property (i.e, the SIRS can rank the important predictors in the top asymptotically). Ji and Jin (2012) proposed the two-stage method: screening by Univariate thresholding and cleaning by Penalized least squares for Selecting variables, namely UPS. They further theoretically demonstrated that under certain settings, the UPS can outperform the LASSO and subset selection, both of which are one-stage approaches. This motivates us to develop more effective screening procedures using two-stage approaches.

In this paper, we propose a new feature screening procedure for ultrahigh dimensional data based on distance correlation. Szekely, Rizzo and Bakirov (2007) and Szekely and Rizzo (2009) showed that the distance correlation of two random vectors equals to zero if and only if these two random vectors are independent. Furthermore, the distance correlation of two univariate normal random variables is a strictly increasing function of the absolute value of the Pearson correlation of these two normal random variables. These two remarkable properties motivate us to use the distance correlation for feature screening in ultrahigh dimensional data. We refer to our Sure Independence Screening procedure based on the Distance Correlation as the DC-SIS. The DC-SIS can be implemented as easily as the SIS. It is equivalent to the SIS when both the response and predictor variables are normally distributed. However, the DC-SIS has appealing features that existing screening procedures including SIS do not possess. For instance, none of the aforementioned screening procedures can handle grouped predictors or multivariate responses. The proposed DC-SIS can be directly employed for screening grouped variables, and it can be directly utilized for ultrahigh dimensional data with multivariate responses. Feature screening for multivariate responses

and/or grouped predictors is of great interest in pathway analyses. As in Chen, et al. (2011), pathway here means sets of proteins that are relevant to specific biological functions without regard to the state of knowledge concerning the interplay among such protein. Since proteins may work interactively to perform various biological functions, pathway analyses complement the marginal association analyses for individual protein, and aim to detect a priori defined set of proteins that are associated with phenotypes of interest. There is a surged interest in pathway analyses in the recent literature (Ashburner, et al., 2000; Mootha, et al., 2003; Subramanian, et al., 2005; Tian, et al., 2005; Bild, et al., 2006; Efron and Tibsirani, 2007; Jones, et al., 2008). Thus, it is of importance to develop feature screening procedures for multivariate responses and/or grouped predictors.

We systematically study the theoretic properties of the DC-SIS, and prove that the DC-SIS possesses the sure screening property in the terminology of Fan and Lv (2008) under very general model settings including linear regression models, for which Fan and Lv (2008) established the sure screening property of the SIS. The sure screening property is a desirable property for feature screening in ultrahigh dimensional data. Even importantly, the DC-SIS can be used for screening features without specifying a regression model between the response and the predictors. Compared with the model-based screening procedures (Fan and Lv, 2008; Fan, Samworth and Wu, 2009; Wang, 2009; Fan and Song, 2010; Fan, Feng and Song, 2011), the DC-SIS is a model-free screening procedure. This virtue makes the proposed procedure robust to model mis-specification. This is a very appealing feature of the proposed procedure in that it may be very difficult in specifying an appropriate regression model for the response and the predictors with little information about the actual model in ultrahigh dimensional data.

We conduct Monte Carlo simulation studies to numerically compare the DC-SIS with the SIS and SIRS. Our simulation results indicate that the DC-SIS can significantly outperform the SIS and the SIRS under many model settings. We also assess the performance of the

DC-SIS as a grouped variable screener, and the simulation results show that the DC-SIS performs very well. We further examine the performance of the DC-SIS for feature screening in ultrahigh dimensional data with multivariate responses; simulation results demonstrate that screening features for multiple responses jointly may have dramatic advantage over screening features with each response separately.

The rest of this paper is organized as follows. In Section 2, we develop the DC-SIS for feature screening and establish its sure screening property. In Section 3, we examine the finite sample performance of the DC-SIS via Monte Carlo simulations. We also illustrate the proposed methodology through a real data example. This paper concludes with a brief discussion in Section 4. All technical proofs are given in the Appendix.

2. INDEPENDENCE SCREENING USING DISTANCE CORRELATION

2.1. Some Preliminaries

Szekely, Rizzo and Bakirov (2007) advocated using the distance correlation for measuring dependence between two random vectors. To be precise, let $\phi_{\mathbf{u}}(\mathbf{t})$ and $\phi_{\mathbf{v}}(\mathbf{s})$ be the respective characteristic functions of the random vectors $\mathbf{u}$ and $\mathbf{v}$, and $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s})$ be the joint characteristic function of $\mathbf{u}$ and $\mathbf{v}$. They defined the distance covariance between $\mathbf{u}$ and $\mathbf{v}$ with finite first moments to be the nonnegative number $\mathrm{dcov}(\mathbf{u},\mathbf{v})$ given by

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = \int_{R^{d_u+d_v}} \|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s}) - \phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})\|^2 \, w(\mathbf{t},\mathbf{s}) \, d\mathbf{t} \, d\mathbf{s}, \tag{2.1}$$

where d_u and d_v are the dimensions of $\mathbf{u}$ and $\mathbf{v}$, respectively, and

$$w(\mathbf{t},\mathbf{s}) = \left\{ c_{d_u} c_{d_v} \|\mathbf{t}\|_{d_u}^{1+d_u} \|\mathbf{s}\|_{d_v}^{1+d_v} \right\}^{-1}$$

with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$. Throughout this paper, $\|\mathbf{a}\|_d$ stands for the Euclidean norm of $\mathbf{a} \in \mathbb{R}^d$, and $\|\phi\|^2 = \phi\bar{\phi}$ for a complex-valued function ϕ with $\bar{\phi}$ being the conjugate of ϕ. The distance correlation (DC) between $\mathbf{u}$ and $\mathbf{v}$ with finite first moments is defined as

$$\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u}, \mathbf{u})\mathrm{dcov}(\mathbf{v}, \mathbf{v})}}. \tag{2.2}$$

Szekely, Rizzo and Bakirov (2007) systematically studied the theoretic properties of the DC.

Two remarkable properties of the DC motivate us to utilize it in a feature screening procedure. The first one is the relationship between the DC and the Pearson correlation coefficient. For two univariate normal random variables U and V with the Pearson correlation coefficient ρ, Szekely, Rizzo and Bakirov (2007) and Szekely and Rizzo (2009) showed that

$$\mathrm{dcorr}(U, V) = \left\{ \frac{\rho \arcsin(\rho) + \sqrt{1-\rho^2} - \rho \arcsin(\rho/2) - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}} \right\}^{1/2}, \tag{2.3}$$

which is strictly increasing in $|\rho|$. This property implies that the DC-based feature screening procedure is equivalent to the marginal Pearson correlation learning for linear regression with normally distributed predictors and random error. In such a situation, Fan and Lv (2008) showed that the Pearson correlation learning has the sure screening property.

The second remarkable property of the DC is $\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are independent (Szekely, Rizzo and Bakirov, 2007). We note that two univariate random variables U and V are independent if and only if U and $T(V)$, a strictly monotone transformation of V, are independent. This implies that a DC-based feature screening procedure can be more effective than the marginal Pearson correlation learning in the presence of nonlinear relationship between U and V. We will demonstrate in the next section that a DC-based screening procedure is a model-free procedure in that one does not need to specify a model structure between the predictors and the response.

Szekely, Rizzo and Bakirov (2007, Remark 3) stated that

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = S_1 + S_2 - 2S_3,$$

where S_j, $j = 1$, 2 and 3, are defined below:

$$\begin{aligned} S_1 &= E\left\{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{d_u}\|\mathbf{v}-\widetilde{\mathbf{v}}\|_{d_v}\right\}, \\ S_2 &= E\left\{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{d_u}\right\} E\left\{\|\mathbf{v}-\widetilde{\mathbf{v}}\|_{d_v}\right\}, \\ S_3 &= E\left\{E\left(\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{d_u}|\ \mathbf{u}\right) E\left(\|\mathbf{v}-\widetilde{\mathbf{v}}\|_{d_v}|\ \mathbf{v}\right)\right\}, \end{aligned} \tag{2.4}$$

where $(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}})$ is an independent copy of $(\mathbf{u},\mathbf{v})$.

Suppose that $\{(\mathbf{u}_i,\mathbf{v}_i), i = 1,\cdots,n\}$ is a random sample from the population $(\mathbf{u},\mathbf{v})$. Szekely, Rizzo and Bakirov (2007) proposed to estimate S_1, S_2 and S_3 through the usual moment estimation. To be precise,

$$\widehat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i-\mathbf{u}_j\|_{d_u}\|\mathbf{v}_i-\mathbf{v}_j\|_{d_v},$$

$$\begin{aligned} \widehat{S}_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i-\mathbf{u}_j\|_{d_u}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{v}_i-\mathbf{v}_j\|_{d_v}, \text{and} \\ \widehat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|\mathbf{u}_i-\mathbf{u}_l\|_{d_u}\|\mathbf{v}_j-\mathbf{v}_l\|_{d_v}. \end{aligned}$$

Thus, a natural estimator of $\mathrm{dcov}^2(\mathbf{u},\mathbf{v})$ is given by

$$\widehat{\mathrm{dcov}}^2(\mathbf{u},\mathbf{v}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3.$$

Similarly, we can define the sample distance covariances $\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{u})$ and $\widehat{\mathrm{dcov}}(\mathbf{v},\mathbf{v})$. Accord-

ingly, the sample distance correlation between $\mathbf{u}$ and $\mathbf{v}$ can be defined by

$$\widehat{\mathrm{dcorr}}(\mathbf{u},\mathbf{v}) = \frac{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{v})}{\sqrt{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{u})\widehat{\mathrm{dcov}}(\mathbf{v},\mathbf{v})}}.$$

2.2. An Independence Ranking and Screening Procedure

In this section we propose an independence screening procedure built upon the DC. Let $\mathbf{y} = (Y_1, \cdots, Y_q)^{\mathrm{T}}$ be the response vector with support Ψ_y, and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be the predictor vector. We regard q as a fixed number in this context. In an ultrahigh-dimensional setting the dimensionality p greatly exceeds the sample size n. It is thus natural to assume that only a small number of predictors are relevant to $\mathbf{y}$. Denote by $F(\mathbf{y} \mid \mathbf{x})$ the conditional distribution function of $\mathbf{y}$ given $\mathbf{x}$. Without specifying a regression model, we define the index set of the active and inactive predictors by

$$\begin{aligned} \mathcal{D} &= \{k : F(\mathbf{y} \mid \mathbf{x}) \text{ functionally depends on } X_k \text{ for some } \mathbf{y} \in \Psi_y\}, \\ \mathcal{I} &= \{k : F(\mathbf{y} \mid \mathbf{x}) \text{ does not functionally depend on } X_k \text{ for any } \mathbf{y} \in \Psi_y\}. \end{aligned} \tag{2.5}$$

We further write $\mathbf{x}_{\mathcal{D}} = \{X_k : k \in \mathcal{D}\}$ and $\mathbf{x}_{\mathcal{I}} = \{X_k : k \in \mathcal{I}\}$, and refer to $\mathbf{x}_{\mathcal{D}}$ as an *active* predictor vector and its complement $\mathbf{x}_{\mathcal{I}}$ as an *inactive* predictor vector. The index subset $\mathcal{D}$ of all active predictors or, equivalently, the index subset $\mathcal{I}$ of all inactive predictors, is the objective of our primary interest. Definition (2.5) implies that $\mathbf{y} \perp\!\!\!\perp \mathbf{x}_{\mathcal{I}} \mid \mathbf{x}_{\mathcal{D}}$, where $\perp\!\!\!\perp$ denotes statistical independence. That is, given $\mathbf{x}_{\mathcal{D}}$, the remaining predictors $\mathbf{x}_{\mathcal{I}}$ are independent of $\mathbf{y}$. Thus the inactive predictors $\mathbf{x}_{\mathcal{I}}$ are redundant when the active predictors $\mathbf{x}_{\mathcal{D}}$ are known.

For ease of presentation, we write

$$\omega_k = \mathrm{dcorr}^2(X_k, \mathbf{y}), \quad \text{and} \quad \widehat{\omega}_k = \widehat{\mathrm{dcorr}}^2(X_k, \mathbf{y}), \text{ for } k = 1, \cdots, p$$

based on a random sample $\{\mathbf{x}_i, \mathbf{y}_i\}$, $i = 1, \ldots, n$. We consider using ω_k as a marginal utility to rank the importance of X_k at the population level. We utilize the DC because it allows for arbitrary regression relationship of $\mathbf{y}$ onto $\mathbf{x}$, regardless of whether it is linear or nonlinear. The DC also permits univariate and multivariate response, regardless of whether it is continuous, discrete or categorical. In addition, it allows for groupwise predictors. Thus, this DC based screening procedure is completely model-free. We select a set of important predictors with large $\hat{\omega}_k$. That is, we define

$$\widehat{\mathcal{D}}^{\star} = \left\{k : \widehat{\omega}_k \geq cn^{-\kappa}, \text{ for } 1 \leq k \leq p\right\},$$

where c and κ are pre-specified threshold values which will be defined in condition (C2) in the subsequent section.

2.3. Theoretical Properties

Next we study the theoretical properties of the proposed independence screening procedure built upon the DC. The following conditions are imposed to facilitate the technical proofs, although they may not be the weakest ones.

(C1) Both $\mathbf{x}$ and $\mathbf{y}$ satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that for all $0 < s \leq 2s_0$,

$$\sup_p \max_{1 \leq k \leq p} E\left\{\exp(s\|X_k\|_1^2)\right\} < \infty, \text{ and } E\{\exp(s\|\mathbf{y}\|_q^2)\} < \infty.$$

(C2) The minimum distance correlation of active predictors satisfies

$$\min_{k \in \mathcal{D}} \omega_k \geq 2cn^{-\kappa}, \text{ for some constants } c > 0 \text{ and } 0 \leq \kappa < 1/2.$$

Condition (C1) follows immediately when $\mathbf{x}$ and $\mathbf{y}$ are bounded uniformly, or when they

have multivariate normal distribution. The normality assumption has been widely used in the area of ultrahigh dimensional data analysis to facilitate the technical derivations. See, for example, Fan and Lv (2008) and Wang (2009).

Next we explore condition (C2). When $\mathbf{x}$ and $\mathbf{y}$ have multivariate normal distribution, (2.3) gives an explicit relationship between the DC and the squared Pearson correlation. For simplicity, we write $\text{dcorr}(X_k, \mathbf{y}) = T_0\left(|\rho(X_k, \mathbf{y})|\right)$ where $T_0(\cdot)$ is strictly increasing given in (2.3). In this situation, condition (C2) requires essentially that $\min_{k\in\mathcal{D}}|\rho(X_k, \mathbf{y})| \geq T_{inv}(2cn^{-\kappa})$, where $T_{inv}(\cdot)$ is the inverse function of $T_0(\cdot)$. This is parallel to condition 3 of Fan and Lv (2008) where it is assumed that $\min_{k\in\mathcal{D}}|\rho(X_k, \mathbf{y})| \geq 2cn^{-\kappa}$. This intuitive illustration implies that condition (C2) requires that the marginal DC of active predictors cannot be too small, which is similar to condition 3 of Fan and Lv (2008). We remark here that, although we illustrate the intuition by assuming that $\mathbf{x}$ and $\mathbf{y}$ are multivariate normal, we do not require this assumption explicitly in our context. The following theorem establishes the sure screening property for the DC-SIS procedure.

Theorem 1. *Under condition (C1), for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants $c_1 > 0$ and $c_2 > 0$ such that*

$$Pr\left(\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\right) \leq O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right). \tag{2.6}$$

Under conditions (C1) and (C2), we have that

$$Pr\left(\mathcal{D} \subseteq \widehat{\mathcal{D}}^{\star}\right) \geq 1 - O\left(s_n\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right), \tag{2.7}$$

where s_n is the cardinality of $\mathcal{D}$.

The sure screening property holds for the DC-SIS under milder conditions than those for the SIS (Fan and Lv, 2008) in that we do not require the regression function of $\mathbf{y}$ onto $\mathbf{x}$

to be linear. Thus, the DC-SIS provides a unified alternative to existing model-based sure screening procedures. Compared with the SIRS, the DC-SIS can effectively handle grouped predictors and multivariate responses.

To balance the two terms in the right hand side of (2.6), we choose the optimal order $\gamma = (1-2\kappa)/3$, then the first part of Theorem 1 becomes

$$\Pr\Big(\max_{1\le k\le p}|\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\Big) \le O\left(p\left[\exp\left\{-c_1 n^{(1-2\kappa)/3}\right\}\right]\right),$$

for some constant $c_1 > 0$, indicating that we can handle the NP-dimensionality of order $\log p = o\left(n^{(1-2\kappa)/3}\right)$. If we further assume that X_k and $\mathbf{y}$ are bounded uniformly in p, then we can obtain without much difficulty that

$$\Pr\left(\max_{1\le k\le p}|\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\right) \le O\left(p\left[\exp\left\{-c_1 n^{1-2\kappa}\right\}\right]\right).$$

In this case, we can handle the NP-dimensionality $\log p = o\left(n^{1-2\kappa}\right)$.

3. NUMERICAL STUDIES

In this section we assess the performance of the DC-SIS by Monte Carlo simulation. Our simulation studies were conducted using R code. We further illustrate the proposed screening procedure with an empirical analysis of a real data example.

In Examples 1, 2 and 3, we generate $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ from normal distribution with zero mean and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$, and the error term ε from standard normal distribution $\mathcal{N}(0,1)$. We consider two covariance matrices to assess the performance of the DC-SIS and to compare with existing methods: (i) $\sigma_{ij} = 0.8^{|i-j|}$ and (ii) $\sigma_{ij} = 0.5^{|i-j|}$. We fix the sample size n to be 200 and vary the dimension p from 2,000 to 5,000. We

repeat each experiment 500 times, and evaluate the performance through the following three criteria.

1. $\mathcal{S}$: the minimum model size to include all active predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of 500 replications.

2. $\mathcal{P}_s$: the proportion that an individual active predictor is selected for a given model size d in the 500 replications.

3. $\mathcal{P}_a$: the proportion that all active predictors are selected for a given model size d in the 500 replications.

The $\mathcal{S}$ is used to measure the model complexity of the resulting model of an underlying screening procedure. The closer to the minimum model size the $\mathcal{S}$ is, the better the screening procedure is. The sure screening property ensures that $\mathcal{P}_s$ and $\mathcal{P}_a$ are both close to one when the estimated model size d is sufficiently large. We choose d to be $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$ throughout our simulations to empirically examine the effect of the cutoff, where $[a]$ denotes the integer part of a.

Example 1. This example is designed to compare the finite sample performance of the DC-SIS with the SIS (Fan and Lv, 2008) and SIRS (Zhu, Li, Li and Zhu, 2011). In this example, we generate the response from the following four models:

$$\begin{aligned}
\textbf{(1.a):} \qquad Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + c_4\beta_4 X_{22} + \varepsilon, \\
\textbf{(1.b):} \qquad Y &= c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) + c_4\beta_3 X_{22} + \varepsilon, \\
\textbf{(1.c):} \qquad Y &= c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) X_{22} + \varepsilon, \\
\textbf{(1.d):} \qquad Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + \exp(c_4|X_{22}|)\varepsilon,
\end{aligned}$$

where $\mathbf{1}(X_{12} < 0)$ is an indicator function. The regression functions $E(Y \mid \mathbf{x})$ in models

Table 1: The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 1.

$\mathcal{S}$	SIS					SIRS					DC-SIS				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
case 1: $p = 2000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$															
(1.a)	4.0	4.0	5.0	7.0	21.2	4.0	4.0	5.0	7.0	45.1	4.0	4.0	4.0	6.0	18.0
(1.b)	68.0	578.5	1180.5	1634.5	1938.0	232.9	871.5	1386.0	1725.2	1942.4	5.0	9.0	24.5	73.0	345.1
(1.c)	395.9	1037.2	1438.0	1745.0	1945.1	238.5	805.0	1320.0	1697.0	1946.0	6.0	10.0	22.0	59.0	324.1
(1.d)	130.5	611.2	1166.0	1637.0	1936.5	42.0	304.2	797.0	1432.2	1846.1	4.0	5.0	9.0	41.0	336.2
case 2: $p = 2000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$															
(1.a)	5.0	9.0	16.0	97.0	729.4	5.0	9.0	18.0	112.8	957.1	4.0	7.0	11.0	31.2	507.2
(1.b)	26.0	283.2	852.0	1541.2	1919.0	103.9	603.0	1174.0	1699.2	1968.0	5.0	8.0	11.0	17.0	98.0
(1.c)	224.5	775.2	1249.5	1670.0	1951.1	118.6	573.2	1201.5	1685.2	1955.0	7.0	10.0	15.0	38.0	198.3
(1.d)	79.0	583.8	1107.5	1626.2	1930.0	50.9	300.5	728.0	1368.2	1900.1	4.0	7.0	17.0	73.2	653.1
case 3: $p = 5000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$															
(1.a)	4.0	4.0	5.0	6.0	59.0	4.0	4.0	5.0	7.0	88.4	4.0	4.0	4.0	6.0	34.1
(1.b)	165.1	1112.5	2729.0	3997.2	4851.5	560.8	1913.0	3249.0	4329.0	4869.1	5.0	11.8	45.0	168.8	956.7
(1.c)	1183.7	2712.0	3604.5	4380.2	4885.0	440.4	1949.0	3205.5	4242.8	4883.1	7.0	17.0	53.0	179.5	732.0
(1.d)	259.9	1338.5	2808.5	3990.8	4764.9	118.7	823.2	1833.5	3314.5	4706.1	4.0	5.0	15.0	77.2	848.2
case 4: $p = 5000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$															
(1.a)	5.0	10.0	26.5	251.5	2522.7	5.0	10.0	28.0	324.8	3246.4	5.0	8.0	14.0	69.0	1455.1
(1.b)	40.7	639.8	2072.0	3803.8	4801.7	215.7	1677.8	3010.0	4352.2	4934.1	5.0	8.0	11.0	21.0	162.0
(1.c)	479.2	1884.8	3347.5	4298.5	4875.2	297.7	1359.2	2738.5	4072.5	4877.6	8.0	12.0	22.0	83.0	657.9
(1.d)	307.0	1544.0	2832.5	4026.2	4785.2	148.2	672.0	1874.0	3330.0	4665.2	4.0	7.0	21.0	165.2	1330.0

(1.a)-**(1.d)** are all nonlinear in X_{12}. In addition, models (**1.b**) and (**1.c**) contain an interaction term X_1X_2, and model (**1.d**) is heteroscedastic. Following Fan and Lv (2008), we choose $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, 2, 3$ and 4, where $a = 4 \log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z \sim \mathcal{N}(0, 1)$. We set $(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$ in this example to challenge the feature screening procedures under consideration. For each independence screening procedure, we compute the associated marginal utility between each predictor X_k and the response Y. That is, we regard $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ as the predictor vector in this example.

Tables 1 and 2 depict the simulation results for $\mathcal{S}$, $\mathcal{P}_s$ and $\mathcal{P}_a$. The performances of the DC-SIS, SIS and SIRS are quite similar in model **(1.a)**, indicating that the SIS has a robust performance if the working linear model does not deviate far from the underlying true model. The DC-SIS outperforms the SIS and SIRS significantly in models **(1.b)**, **(1.c)** and **(1.d)**. Both the SIS and SIRS have little chance to identify the important predictors X_1 and X_2 in models **(1.b)** and **(1.c)**, and X_{22} in model **(1.d)**.

Example 2. We illustrate that the DC-SIS can be directly used for screening grouped

Table 2: The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 1. The user-specified model sizes $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

		SIS					SIRS					DC-SIS				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
model	size	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
case 1: $p = 2000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$																
	d_1	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.95	1.00	0.94	1.00	1.00	0.97	1.00	0.96
(1.a)	d_2	1.00	1.00	0.98	1.00	0.97	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.98	1.00	0.98
	d_3	1.00	1.00	0.98	1.00	0.98	1.00	1.00	0.97	1.00	0.97	1.00	1.00	0.99	1.00	0.98
	d_1	0.08	0.07	0.97	1.00	0.03	0.02	0.03	0.98	1.00	0.00	0.72	0.70	0.99	1.00	0.58
(1.b)	d_2	0.12	0.13	0.98	1.00	0.06	0.05	0.05	0.99	1.00	0.01	0.85	0.84	1.00	1.00	0.76
	d_3	0.15	0.17	0.99	1.00	0.07	0.06	0.06	0.99	1.00	0.01	0.89	0.88	1.00	1.00	0.82
	d_1	0.12	0.13	0.01	0.99	0.00	0.04	0.03	0.51	1.00	0.01	0.93	0.93	0.77	1.00	0.65
(1.c)	d_2	0.17	0.18	0.03	0.99	0.00	0.07	0.05	0.67	1.00	0.01	0.97	0.96	0.84	1.00	0.79
	d_3	0.21	0.21	0.05	0.99	0.00	0.09	0.08	0.75	1.00	0.02	0.98	0.97	0.89	1.00	0.84
	d_1	0.42	0.22	0.14	0.42	0.02	1.00	0.98	0.87	0.05	0.04	1.00	0.91	0.81	0.99	0.73
(1.d)	d_2	0.48	0.29	0.22	0.50	0.03	1.00	0.99	0.91	0.10	0.09	1.00	0.94	0.87	1.00	0.82
	d_3	0.56	0.32	0.26	0.54	0.04	1.00	0.99	0.93	0.12	0.11	1.00	0.96	0.92	1.00	0.88
case 2: $p = 2000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$																
	d_1	1.00	1.00	0.63	1.00	0.63	1.00	1.00	0.62	1.00	0.62	1.00	1.00	0.78	1.00	0.77
(1.a)	d_2	1.00	1.00	0.71	1.00	0.72	1.00	1.00	0.70	1.00	0.69	1.00	1.00	0.84	1.00	0.84
	d_3	1.00	1.00	0.77	1.00	0.78	1.00	1.00	0.75	1.00	0.75	1.00	1.00	0.86	1.00	0.86
	d_1	0.12	0.13	0.81	1.00	0.06	0.04	0.04	0.88	1.00	0.02	0.97	0.98	0.92	1.00	0.88
(1.b)	d_2	0.19	0.19	0.86	1.00	0.12	0.07	0.07	0.91	1.00	0.03	0.99	0.99	0.95	1.00	0.94
	d_3	0.22	0.23	0.88	1.00	0.15	0.09	0.11	0.93	1.00	0.06	1.00	0.99	0.96	1.00	0.96
	d_1	0.17	0.16	0.03	0.99	0.00	0.04	0.04	0.53	1.00	0.02	1.00	1.00	0.75	1.00	0.75
(1.c)	d_2	0.22	0.22	0.06	1.00	0.01	0.08	0.08	0.71	1.00	0.03	1.00	1.00	0.85	1.00	0.86
	d_3	0.27	0.27	0.10	1.00	0.03	0.10	0.10	0.81	1.00	0.05	1.00	1.00	0.90	1.00	0.90
	d_1	0.44	0.38	0.11	0.45	0.03	1.00	1.00	0.73	0.05	0.04	0.99	0.98	0.68	1.00	0.67
(1.d)	d_2	0.51	0.46	0.18	0.53	0.05	1.00	1.00	0.81	0.09	0.08	1.00	0.98	0.76	1.00	0.75
	d_3	0.55	0.49	0.22	0.57	0.06	1.00	1.00	0.84	0.14	0.11	1.00	0.99	0.80	1.00	0.80
case 3: $p = 5000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$																
	d_1	1.00	1.00	0.94	1.00	0.94	1.00	0.99	0.92	1.00	0.92	1.00	0.99	0.96	1.00	0.95
(1.a)	d_2	1.00	1.00	0.95	1.00	0.95	1.00	1.00	0.95	1.00	0.95	1.00	1.00	0.97	1.00	0.97
	d_3	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.98	1.00	0.98
	d_1	0.06	0.06	0.94	1.00	0.02	0.02	0.02	0.96	1.00	0.00	0.59	0.60	0.98	1.00	0.46
(1.b)	d_2	0.09	0.09	0.96	1.00	0.03	0.03	0.03	0.97	1.00	0.01	0.72	0.72	0.99	1.00	0.61
	d_3	0.12	0.10	0.97	1.00	0.04	0.05	0.04	0.98	1.00	0.01	0.79	0.78	0.99	1.00	0.68
	d_1	0.06	0.06	0.01	0.99	0.00	0.03	0.02	0.30	1.00	0.00	0.86	0.87	0.61	1.00	0.41
(1.c)	d_2	0.10	0.10	0.02	1.00	0.00	0.04	0.03	0.45	1.00	0.00	0.92	0.93	0.69	1.00	0.57
	d_3	0.12	0.12	0.02	1.00	0.00	0.05	0.05	0.53	1.00	0.00	0.94	0.95	0.73	1.00	0.64
	d_1	0.39	0.21	0.11	0.40	0.01	1.00	0.97	0.82	0.02	0.02	0.99	0.87	0.74	0.99	0.65
(1.d)	d_2	0.44	0.24	0.14	0.45	0.01	1.00	0.98	0.88	0.04	0.03	0.99	0.90	0.81	0.99	0.75
	d_3	0.48	0.28	0.17	0.47	0.02	1.00	0.99	0.90	0.06	0.05	0.99	0.92	0.85	1.00	0.79
case 4: $p = 5000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$																
	d_1	1.00	1.00	0.55	1.00	0.55	1.00	1.00	0.55	1.00	0.55	1.00	1.00	0.70	1.00	0.69
(1.a)	d_2	1.00	1.00	0.61	1.00	0.62	1.00	1.00	0.61	1.00	0.61	1.00	1.00	0.76	1.00	0.76
	d_3	1.00	1.00	0.67	1.00	0.67	1.00	1.00	0.64	1.00	0.64	1.00	1.00	0.80	1.00	0.80
	d_1	0.10	0.09	0.74	1.00	0.05	0.02	0.02	0.83	1.00	0.00	0.94	0.94	0.90	1.00	0.82
(1.b)	d_2	0.12	0.13	0.81	1.00	0.07	0.03	0.04	0.87	1.00	0.01	0.97	0.97	0.93	1.00	0.89
	d_3	0.15	0.16	0.84	1.00	0.10	0.05	0.06	0.90	1.00	0.02	0.98	0.98	0.95	1.00	0.92
	d_1	0.10	0.10	0.02	0.98	0.00	0.02	0.03	0.34	1.00	0.00	1.00	1.00	0.64	1.00	0.63
(1.c)	d_2	0.13	0.14	0.04	0.99	0.01	0.04	0.04	0.50	1.00	0.01	1.00	1.00	0.74	1.00	0.74
	d_3	0.16	0.18	0.05	0.99	0.01	0.05	0.05	0.61	1.00	0.02	1.00	1.00	0.79	1.00	0.79
	d_1	0.42	0.32	0.09	0.40	0.01	1.00	1.00	0.66	0.02	0.01	0.99	0.97	0.63	0.98	0.59
(1.d)	d_2	0.48	0.39	0.12	0.44	0.02	1.00	1.00	0.74	0.04	0.03	0.99	0.97	0.70	1.00	0.68
	d_3	0.51	0.42	0.15	0.46	0.02	1.00	1.00	0.78	0.05	0.04	0.99	0.98	0.73	1.00	0.71

predictors. In many regression problems, some predictors can be naturally grouped. The most common example which contains group variables is the multi-factor ANOVA problem, in which each factor may have several levels and can be expressed through a group of dummy variables. The goal of ANOVA is to select important main effects and interactions for accurate predictions, which amounts to the selection of groups of dummy variables. To demonstrate the practicability of the DC-SIS, we adopt the following model:

$$\begin{aligned} Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3\{\mathbf{1}(X_{12} < q_1) + 1.5 \times \mathbf{1}(q_1 \le X_{12} < q_2) \\ &\quad +2 \times \mathbf{1}(q_2 \le X_{12} < q_3)\} + c_4\beta_4 X_{22} + \varepsilon, \end{aligned}$$

where q_1, q_2 and q_3 are the 25%, 50% and 75% quantiles of X_{12}, respectively. The variables X with the coefficients c_i's and β_i's are the same as those in Example 1. We write

$$\widetilde{\mathbf{x}}_{12} = \{\mathbf{1}(X_{12} < q_1),\ \mathbf{1}(q_1 \le X_{12} < q_2),\ \mathbf{1}(q_2 \le X_{12} < q_3))\}^{\mathrm{T}}.$$

These three correlated variables naturally become a group. The predictor vector in this example becomes $\mathbf{x} = (X_1, \ldots, X_{11}, \widetilde{\mathbf{x}}_{12}, X_{13}, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^{p+2}$. We remark here that the marginal utility of the grouped variable $\widetilde{\mathbf{x}}_{12}$ is defined by

$$\widehat{\omega}_{12} = \widehat{\mathrm{dcorr}}^2(\widetilde{\mathbf{x}}_{12}, Y).$$

The 5%, 25%, 50%, 75% and 95% percentiles of the minimum model size $\mathcal{S}$ are summarized in Table 3. These percentiles indicate that with very high probability, the minimum model size $\mathcal{S}$ to ensure the inclusion of all active predictors is small. Note that $[n/\log(n)] = 37$. Thus, almost all $\mathcal{P}_s$s and $\mathcal{P}_a$s equal 100%. All active predictors including the grouped variable $\widetilde{\mathbf{x}}_{12}$ can almost perfectly be selected into the resulting model across all three different model sizes. Hence, the DC-SIS is efficient to select the grouped predictors.

Table 3: The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 2.

$\mathcal{S}$	$p = 2000$					$p = 5000$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$\sigma_{ij} = 0.5^{\|i-j\|}$	4.0	4.0	4.0	5.0	12.0	4.0	4.0	4.0	6.0	16.1
$\sigma_{ij} = 0.8^{\|i-j\|}$	4.0	5.0	7.0	9.0	15.2	4.0	5.0	7.0	9.0	21.0

Example 3. In this example, we investigate the performance of the DC-SIS with multivariate responses. The SIS proposed in Fan and Lv (2008) cannot be directly applied for such settings. In contrast, the DC-SIS is ready for screening the active predictors by the nature of DC. In this example, we generate $\mathbf{y} = (Y_1, Y_2)^{\mathrm{T}}$ from normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}} = (\sigma_{\mathbf{x},ij})_{2\times 2}$, where $\sigma_{\mathbf{x},11} = \sigma_{\mathbf{x},22} = 1$ and $\sigma_{\mathbf{x},12} = \sigma_{\mathbf{x},21} = \sigma(\mathbf{x})$. We consider two scenarios for the correlation function $\sigma(\mathbf{x})$:

(3.a): $\sigma(\mathbf{x}) = \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, where $\boldsymbol{\beta}_1 = (0.8, 0.6, 0, \ldots, 0)^{\mathrm{T}}$.

(3.b): $\sigma(\mathbf{x}) = \{\exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) - 1\} / \{\exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) + 1\}$, where $\boldsymbol{\beta}_2 = (2 - U_1, 2 - U_2, 2 - U_3, 2 - U_4, 0, \ldots, 0)^{\mathrm{T}}$ with U_i's being independent and identically distributed according to uniform distribution Uniform$[0, 1]$.

Tables 4 and 5 depict the simulation results. Table 4 implies that the DC-SIS performs reasonably well for both models (**3.a**) and (**3.b**) in terms of model complexity. Table 5 indicates that the proportions that the active predictors are selected into the model are close to one, which supports the assertion that the DC-SIS processes the sure screening property. It implies that the DC-SIS can identify the active predictors contained in correlations between multivariate responses. This may be potentially useful in gene co-expression analysis.

Example 4. The Cardiomyopathy microarray dataset was once analyzed by Segal, Dahlquist and Conklin (2003) and Hall and Miller (2009). The goal is to identify the most influential genes for overexpression of a G protein-coupled receptor (Ro1) in mice. The response Y is the Ro1 expression level, and the predictors X_k's are other gene expression levels. Compared

Table 4: The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 3.

	$\mathcal{S}$	$p = 2000$					$p = 5000$				
	Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$\sigma_{ij} = 0.5^{\|i-j\|}$	**(3.a)**	4.0	9.0	18.0	39.3	112.3	6.0	22.0	48.0	95.3	296.4
	(3.b)	6.0	19.0	43.0	92.0	253.1	14.0	45.0	92.5	198.8	571.6
$\sigma_{ij} = 0.8^{\|i-j\|}$	**(3.a)**	2.0	3.0	6.0	12.0	40.0	2.0	6.0	14.0	32.0	98.0
	(3.b)	4.0	4.0	4.0	6.0	10.0	4.0	4.0	5.0	8.0	18.1

Table 5: The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 3. The user-specified model sizes $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

		$p = 2000$								$p = 5000$							
		(3.a)			**(3.b)**					**(3.a)**			**(3.b)**				
		$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
	size	X_1	X_2	ALL	X_1	X_2	X_3	X_4	ALL	X_1	X_2	ALL	X_1	X_2	X_3	X_4	ALL
$\sigma_{ij} = 0.5^{\|i-j\|}$	d_1	0.95	0.76	0.74	0.71	0.98	0.98	0.72	0.47	0.79	0.49	0.42	0.48	0.91	0.90	0.53	0.20
	d_2	0.98	0.90	0.90	0.85	0.99	0.99	0.85	0.71	0.93	0.70	0.67	0.67	0.97	0.97	0.71	0.45
	d_3	1.00	0.95	0.95	0.91	0.99	1.00	0.90	0.81	0.97	0.81	0.80	0.75	0.98	0.99	0.78	0.55
$\sigma_{ij} = 0.8^{\|i-j\|}$	d_1	0.98	0.95	0.94	1.00	1.00	1.00	1.00	1.00	0.92	0.84	0.81	1.00	1.00	1.00	0.99	0.99
	d_2	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.93	1.00	1.00	1.00	1.00	1.00
	d_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.96	1.00	1.00	1.00	1.00	1.00

with the sample size $n = 30$ in this dataset, the dimension $p = 6319$ is very large.

The DC-SIS procedure ranks two genes, labeled Msa.2134.0 and Msa.2877.0, at the top. The scatter plots of Y versus these two gene expression levels with cubic spline fit curves in Figure 1 indicate clearly the existence of nonlinear patterns. Yet, our finding is different from Hall and Miller (2009) in that they ranked Msa.2877.0 and Msa.1166.0 at the top with their proposed generalized correlation ranking. A natural question arises: which screening procedure performs better in terms of ranking? To compare the performance of these two procedures, we fit an additive model as follows:

$$Y = \ell_{k1}(X_{k1}) + \ell_{k2}(X_{k2}) + \varepsilon_k, \text{ for } k = 1, 2.$$

The DC-SIS, corresponding to $k = 1$, regards Msa.2134.0 and Msa.2877.0 as the two predictors, while the generalized correlation ranking proposed by Hall and Miller (2009), corre-

sponding to $k = 2$, regards Msa.2877.0 and Msa.1166.0 as predictors in the above model. We fit the unknown link functions ℓ_{ki} using the R `mgcv` package. The DC-SIS method clearly achieves better performance with the adjusted R^2 of 96.8% and the deviance explained of 98.3%, in contrast to the adjusted R^2 of 84.5% and the deviance explained of 86.6% for the generalized correlation ranking method. We remark here that deviance explained means the proportion of the null deviance explained by the proposed model, with a larger value indicating better performance. Because both the adjusted R^2 values and the explained deviance are very large, it seems unnecessary to extract any additional genes.

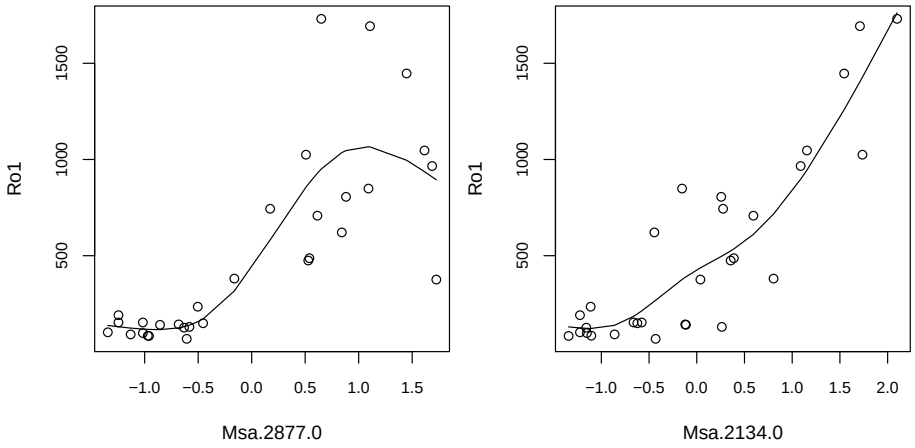

Figure 1. The scatter plot of Y versus two genes expression levels identified by the DC-SIS.

4. DISCUSSION

In this paper we proposed a sure independence screening procedure using distance correlation. We established the sure screening property for this procedure when the number of predictors diverges with an exponential rate of the sample size. We examined the finite-sample performance of the proposed procedure via Monte Carlo studies and illustrated the proposed methodology through a real data example. We followed Fan and Lv (2008) to set the cutoff d in this paper and examine the effect of different values of d. As pointed out by a referee, the choice of d is very important in the screening stage. Zhao and Li (2012)

proposed an approach to selecting d for Cox models based on controlling false positive rate. Their approach is merely for model-based feature screening methods. Zhu, Li, Li and Zhu (2011) proposed an alternative method to determine d for the SIRS. One may adopt their procedure for the DC-SIS. We opt not to pursue this further. Certainly, the selection of d is similar to selection of the tuning parameter in regularization methods, and plays an important role in practical implementation. This is a good topic for future research.

Similar to the SIS, the DC-SIS may fail to identify some important predictors which are marginally independent of the response. Thus, it is of interest to develop an iterative procedure to fix such an issue. In the earlier version of this paper, we proposed an iterative version of DC-SIS. Our empirical studies including Monte Carlo simulation and real data analysis imply that the proposed iterative DC-SIS may be used to fix the problem in a similar spirit of ISIS (Fan and Lv, 2008). Theoretical analysis of the iterative DC-SIS needs further study. New methods to deal with identification of important predictors which are marginally independent of the response is an important topic for future research.

APPENDIX

Appendix A: Some Lemmas

Lemmas 1 and 2 will be used repeatedly in the proof of Theorem 1. These two lemmas provide us two exponential inequalities, and are extracted from Lemma 5.6.1.A and Theorem 5.6.1.A of Serfling (1980, page 200-201).

Lemma 1. *Let* $\mu = E(Y)$. *If* $\Pr(a \leq Y \leq b) = 1$, *then*

$$E\left[\exp\left\{s(Y-\mu)\right\}\right] \leq \exp\left\{s^2(b-a)^2/8\right\}, \text{ for any } s > 0.$$

Lemma 2. *Let* $h(Y_1, \cdots, Y_m)$ *be a kernel of the U-statistic* U_n, *and* $\theta = E\left\{h(Y_1, \cdots, Y_m)\right\}$.

If $a \leq h(Y_1, \cdots, Y_m) \leq b$*, then, for any* $t > 0$ *and* $n \geq m$*,*

$$Pr\left(U_n - \theta \geq t\right) \leq \exp\left\{-2[n/m]t^2/(b-a)^2\right\},$$

where $[n/m]$ *denotes the integer part of* n/m*.*

Due to the symmetry of U-statistic, Lemma 2 entails that

$$\Pr\left(|U_n - \theta| \geq t\right) \leq 2\exp\left\{-2[n/m]t^2/(b-a)^2\right\}.$$

Let us introduce some notations before giving the proof of Theorem 1. Let $\{\tilde{X}_k, \tilde{\mathbf{y}}\}$ be an independent copy of $\{X_k, \mathbf{y}\}$, and define $S_{k1} = E\|X_k - \tilde{X}_k\|_1\|\mathbf{y} - \tilde{\mathbf{y}}\|_q$, $S_{k2} = E\|X_k - \tilde{X}_k\|_1 E\|\mathbf{y} - \tilde{\mathbf{y}}\|_q$, and $S_{k3} = E\{E(\|X_k - \tilde{X}_k\|_1|X_k)E(\|\mathbf{y} - \tilde{\mathbf{y}}\|_q|\mathbf{y})\}$, and their sample counterparts

$$\begin{aligned}
\widehat{S}_{k1} &= \frac{1}{n^2}\sum_{i,j=1}^{n}\|X_{ik} - X_{jk}\|_1\|\mathbf{y}_i - \mathbf{y}_j\|_q, \\
\widehat{S}_{k2} &= \frac{1}{n^2}\sum_{i,j=1}^{n}\|X_{ik} - X_{jk}\|_1 \frac{1}{n^2}\sum_{i,j=1}^{n}\|\mathbf{y}_i - \mathbf{y}_j\|_q, \\
\widehat{S}_{k3} &= \frac{1}{n^3}\sum_{i,j,l=1}^{n}\|X_{ik} - X_{lk}\|_1\|\mathbf{y}_j - \mathbf{y}_l\|_q.
\end{aligned}$$

By definitions of distance covariance and sample distance covariance, it follows that

$$\mathrm{dcov}^2(X_k, \mathbf{y}) = S_{k1} + S_{k2} - 2S_{k3} \quad \text{and} \quad \widehat{\mathrm{dcov}}^2(X_k, \mathbf{y}) = \widehat{S}_{k1} + \widehat{S}_{k2} - 2\widehat{S}_{k3}.$$

Appendix B: Proof of Theorem 1

We aim to show the uniform consistency of the denominator and the numerator of $\widehat{\omega}_k$ under regularity conditions respectively. Because the denominator of $\widehat{\omega}_k$ has a similar form

as the numerator, we deal with its numerator only below. Throughout proof, the notations C and c are generic constants which may take different values at each appearance.

We first deal with $\widehat{S}_{k1}$. Define $\widehat{S}^*_{k1} = \{n(n-1)\}^{-1}\sum_{i\neq j}\|X_{ik}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q$, which is a usual U-statistic. We shall establish the uniform consistency of $\widehat{S}^*_{k1}$ by using the theory of U-statistics (Serfling, 1980, Section 5). By using the Cauchy-Schwartz inequality,

$$\begin{aligned} S_{k1} &= E\left(\|X_{ik}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q\right) \leq \left\{E\left(\|X_{ik}-X_{jk}\|_1^2\right)E\left(\|\mathbf{y}_i-\mathbf{y}_j\|_q^2\right)\right\}^{1/2} \\ &\leq 4\left\{E(X_k^2)E\|\mathbf{y}\|_q^2\right\}^{1/2}. \end{aligned}$$

This together with condition (C1) implies that S_{k1} is uniformly bounded in p, that is, $\sup_p \max_{1\leq k\leq p} S_{k1} < \infty$. For any given $\varepsilon > 0$, take n large enough such that $S_{k1}/n < \varepsilon$. Then it can be easily shown that

$$\begin{aligned} \Pr\left(\left|\widehat{S}_{k1}-S_{k1}\right|\geq 2\varepsilon\right) &= \Pr\left\{\left|\widehat{S}^*_{k1}(n-1)/n - S_{k1}(n-1)/n - S_{k1}/n\right|\geq 2\varepsilon\right\} \\ &\leq \Pr\left\{\left|\widehat{S}^*_{k1}-S_{k1}\right|(n-1)/n \geq 2\varepsilon - S_{k1}/n\right\} \\ &\leq \Pr\left(\left|\widehat{S}^*_{k1}-S_{k1}\right|\geq \varepsilon\right). \end{aligned} \tag{B.1}$$

To establish the uniform consistency of $\widehat{S}_{k1}$, it thus suffices to show the uniform consistency of $\widehat{S}^*_{k1}$. Let $h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j) = \|X_{ik}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q$ be the kernel of the U-statistic $\widehat{S}^*_{k1}$. We decompose the kernel function h_1 into two parts: $h_1 = h_1\mathbf{1}(h_1 > M) + h_1\mathbf{1}(h_1 \leq M)$ where M will be specified later. The U-statistic can now be written as follows,

$$\begin{aligned} \widehat{S}^*_{k1} &= \{n(n-1)\}^{-1}\sum_{i\neq j} h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\leq M\right\} \\ &+ \{n(n-1)\}^{-1}\sum_{i\neq j} h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)> M\right\} \\ &= \widehat{S}^*_{k1,1}+\widehat{S}^*_{k1,2}. \end{aligned}$$

Accordingly, we decompose S_{k1} into two parts:

$$\begin{aligned} S_{k1} &= E\left[h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\le M\right\}\right] \\ &+ E\left[h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)> M\right\}\right] \\ &= S_{k1,1}+S_{k1,2}. \end{aligned}$$

Clearly, $\widehat{S}^*_{k1,1}$ and $\widehat{S}^*_{k1,2}$ are unbiased estimators of $S_{k1,1}$ and $S_{k1,2}$, respectively.

We deal with the consistency of $\widehat{S}^*_{k1,1}$ first. With the Markov's inequality, for any $t>0$, we can obtain that

$$\Pr(\widehat{S}^*_{k1,1}-S_{k1,1}\ge\varepsilon)\le\exp\left(-t\varepsilon\right)\exp(-tS_{k1,1})E\{\exp(t\widehat{S}^*_{k1,1})\}.$$

Serfling (1980, Section 5.1.6) showed that any U-statistic can be represented as an average of averages of independent and identically distributed (i.i.d) random variables. That is, $\widehat{S}^*_{k1,1}=(n!)^{-1}\sum_{n!}\Omega_1(X_{1k},\mathbf{y}_1;\cdots;X_{nk},\mathbf{y}_n)$, where $\sum_{n!}$ denotes the summation over all possible permutations of $(1,\ldots,n)$, and each $\Omega_1(X_{1k},\mathbf{y}_1;\cdots;X_{nk},\mathbf{y}_n)$ is an average of $m=[n/2]$ i.i.d random variables (i.e., $\Omega_1=m^{-1}\sum_r h_1^{(r)}\mathbf{1}\{h_1^{(r)}\le M\}$). Since the exponential function is convex, it follows from Jensen's inequality that, for $0<t\le 2s_0$,

$$\begin{aligned} E\{\exp(t\widehat{S}^*_{k1,1})\} &= E\big[\exp\big\{t(n!)^{-1}\sum_{n!}\Omega_1(X_{1k},\mathbf{y}_1;\cdots;X_{nk},\mathbf{y}_n)\big\}\big] \\ &\le (n!)^{-1}\sum_{n!}E\left[\exp\left\{t\Omega_1(X_{1k},\mathbf{y}_1;\cdots;X_{nk},\mathbf{y}_n)\right\}\right] \\ &= E^m\big\{\exp\big(m^{-1}th_1^{(r)}\mathbf{1}\{h_1^{(r)}\le M\}\big)\big\}, \end{aligned}$$

which together with Lemma 1 entails immediately that

$$\begin{aligned}\Pr(\widehat{S}^*_{k1,1} - S_{k1,1} \geq \varepsilon) &\leq \exp(-t\varepsilon)\, E^m\big\{\exp\big(m^{-1}t\big[h_1^{(r)}\mathbf{1}\{h_1^{(r)} \leq M\} - S_{k1,1}\big]\big)\big\} \\ &\leq \exp\big\{-t\varepsilon + M^2t^2/(8m)\big\}.\end{aligned}$$

By choosing $t = 4\varepsilon m/M^2$, we have $\Pr(\widehat{S}^*_{k1,1} - S_{k1,1} \geq \varepsilon) \leq \exp(-2\varepsilon^2 m/M^2)$. Therefore, by the symmetry of U-statistic, we can obtain easily that

$$\Pr\big(\big|\widehat{S}^*_{k1,1} - S_{k1,1}\big| \geq \varepsilon\big) \leq 2\exp\big(-2\varepsilon^2 m/M^2\big). \tag{B.2}$$

Next we show the consistency of $\widehat{S}^*_{k1,2}$. With Cauchy-Schwartz and Markov's inequality,

$$\begin{aligned}S^2_{k1,2} &\leq E\big\{h_1^2(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\big\} \Pr\{h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) > M\} \\ &\leq E\big\{h_1^2(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\big\} E\left[\exp\{s'h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\}\right] / \exp(s'M),\end{aligned}$$

for any $s' > 0$. Using the fact $(a^2 + b^2)/2 \geq (a+b)^2/4 \geq |ab|$, we have

$$\begin{aligned}& h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) = \big\{(X_{ik} - X_{jk})^2(\mathbf{y}_i - \mathbf{y}_j)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{y}_j)\big\}^{1/2} \\ \leq\ & 2\big\{\big(X_{ik}^2 + X_{jk}^2\big)\big(\|\mathbf{y}_i\|_q^2 + \|\mathbf{y}_j\|_q^2\big)\big\}^{1/2} \leq \Big\{\big(X_{ik}^2 + X_{jk}^2 + \|\mathbf{y}_i\|_q^2 + \|\mathbf{y}_j\|_q^2\big)^2\Big\}^{1/2} \\ =\ & X_{ik}^2 + X_{jk}^2 + \|\mathbf{y}_i\|_q^2 + \|\mathbf{y}_j\|_q^2,\end{aligned}$$

which yields that

$$\begin{aligned}E\left[\exp\{s'h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\}\right] &\leq E\left[\exp\big\{s'\big(X_{ik}^2 + X_{jk}^2 + \|\mathbf{y}_i\|_q^2 + \|\mathbf{y}_j\|_q^2\big)\big\}\right] \\ &\leq E\big\{\exp\big(2s'X_{ik}^2\big)\big\}\, E\big\{\exp\big(2s'\|\mathbf{y}_i\|_q^2\big)\big\}.\end{aligned}$$

The last inequality follows from the Cauchy-Schwartz inequality. If we choose $M = cn^\gamma$ for

$0 < \gamma < 1/2 - \kappa$, then $S_{k1,2} \leq \varepsilon/2$ when n is sufficiently large. Consequently,

$$\Pr\big(\big|\widehat{S}^*_{k1,2} - S_{k1,2}\big| > \varepsilon\big) \;\; \leq \;\; \Pr\big(\big|\widehat{S}^*_{k1,2}\big| > \varepsilon/2\big). \tag{B.3}$$

It remains to bound the probability $\Pr\big(\big|\widehat{S}^*_{k1,2}\big| > \varepsilon/2\big)$. We observe that the events satisfy

$$\big\{\big|\widehat{S}^*_{k1,2}\big| > \varepsilon/2\big\} \subseteq \big\{X_{ik}^2 + \|\mathbf{y}_i\|_q^2 > M/2, \text{ for some } 1 \leq i \leq p\big\}. \tag{B.4}$$

To see this, we assume that $X_{ik}^2 + \|\mathbf{y}_i\|_q^2 \leq M/2$ for all $1 \leq i \leq p$. This assumption will lead to a contradiction. To be precise, under this assumption, $h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) \leq X_{ik}^2 + X_{jk}^2 + \|\mathbf{y}_i\|_q^2 + \|\mathbf{y}_j\|_q^2 \leq M$. Consequently, $\big|\widehat{S}^*_{k1,2}\big| = 0$, which is a contrary to the event $\big|\widehat{S}^*_{k1,2}\big| > \varepsilon/2$. This verifies the relation (B.4) is true.

By invoking condition (C1), there must exist a constant C such that

$$\Pr(\|X_k\|_1^2 + \|\mathbf{y}\|_q^2 \geq M/2) \leq \Pr(\|X_k\|_1 \geq \sqrt{M}/2) + \Pr(\|\mathbf{y}\|_q \geq \sqrt{M}/2) \leq 2C\exp(-sM/4).$$

The last inequality follows from Markov's inequality for $s > 0$. Consequently,

$$\begin{aligned} \max_{1\leq k\leq p} \Pr\big(\big|\widehat{S}^*_{k1,2}\big| > \varepsilon/2\big) \;\; &\leq \;\; n \max_{1\leq k\leq p} \Pr(\|X_k\|_1^2 + \|\mathbf{y}\|_q^2 \geq M/2) \\ &\leq \;\; 2nC\exp(-sM/4). \end{aligned} \tag{B.5}$$

Recall that $M = cn^\gamma$. Combining the results (B.2), (B.3) and (B.5), we have

$$\Pr\big(\big|\widehat{S}_{k1} - S_{k1}\big| \geq 4\varepsilon\big) \leq 2\exp\left(-\varepsilon^2 n^{1-2\gamma}\right) + 2nC\exp\left(-sn^\gamma/4\right). \tag{B.6}$$

In the sequel we turn to $\widehat{S}_{k2}$. We write $\widehat{S}_{k2} = \widehat{S}_{k2,1}\widehat{S}_{k2,2}$, where $\widehat{S}_{k2,1} = n^{-2}\sum\limits_{i\neq j}\|X_{ik} - X_{jk}\|_1$, and $\widehat{S}_{k2,2} = n^{-2}\sum\limits_{i\neq j}\|\mathbf{y}_i - \mathbf{y}_j\|_q$. Similarly, we write $S_{k2} = S_{k2,1}S_{k2,2}$, where $S_{k2,1} = E\{\|X_{ik} -$

$X_{jk}\|_1\}$ and $S_{k2,2} = E\{\|\mathbf{y}_i - \mathbf{y}_j\|_q\}$. Following arguments for proving (B.6) we can show that

$$\begin{aligned} &\Pr\big(\big|\widehat{S}_{k2,1} - S_{k2,1}\big| \geq 4\varepsilon\big) \leq 2\exp\left(-\varepsilon^2 n^{1-2\gamma}\right) + 2nC\exp\left(-sn^{2\gamma}/4\right), \text{ and} \\ &\Pr\big(\big|\widehat{S}_{k2,2} - S_{k2,2}\big| \geq 4\varepsilon\big) \leq 2\exp\left(-\varepsilon^2 n^{1-2\gamma}\right) + 2nC\exp\left(-sn^{2\gamma}/4\right). \end{aligned} \tag{B.7}$$

Condition (C1) ensures that $S_{k2,1} \leq \{E(\|X_{ik} - X_{jk}\|_1^2)\}^{1/2} \leq \{4E(X_k^2)\}^{1/2}$ and $S_{k2,2} \leq \big\{E(\|\mathbf{y}_i - \mathbf{y}_j\|_q^2)\big\}^{1/2} \leq \big\{4E(\|\mathbf{y}\|_q^2)\big\}^{1/2}$ are uniformly bounded. That is,

$$\max\Big\{\max_{1\leq k\leq p} S_{k2,1}, S_{k2,2}\Big\} \leq C,$$

for some constant C. Using (B.7) repetitively, we can easily prove that

$$\begin{aligned} \Pr\big\{\big|(\widehat{S}_{k2,1} - S_{k2,1})S_{k2,2}\big| \geq \varepsilon\big\} &\leq \Pr\big(\big|\widehat{S}_{k2,1} - S_{k2,1}\big| \geq \varepsilon/C\big) \\ &\leq 2\exp\big\{-\varepsilon^2 n^{1-2\gamma}/(16C^2)\big\} + 2nC\exp\left(-sn^{2\gamma}/4\right), \\ \Pr\big(\big|S_{k2,1}(\widehat{S}_{k2,2} - S_{k2,2})\big| \geq \varepsilon\big) &\leq \Pr\big(\big|\widehat{S}_{k2,2} - S_{k2,2}\big| \geq \varepsilon/C\big) \\ &\leq 2\exp\big\{-\varepsilon^2 n^{1-2\gamma}/(16C^2)\big\} + 2nC\exp\left(-sn^{2\gamma}/4\right), \end{aligned} \tag{B.8}$$

and

$$\begin{aligned} &\Pr\big\{\big|(\widehat{S}_{k2,1} - S_{k2,1})(\widehat{S}_{k2,2} - S_{k2,2})\big| \geq \varepsilon\big\} \\ \leq &\Pr\big(\big|\widehat{S}_{k2,1} - S_{k2,1}\big| \geq \sqrt{\varepsilon}\big) + \Pr\big(\big|\widehat{S}_{k2,2} - S_{k2,2}\big| \geq \sqrt{\varepsilon}\big) \\ \leq &4\exp\left(-\varepsilon n^{1-2\gamma}/16\right) + 4nC\exp\left(-sn^{2\gamma}/4\right). \end{aligned} \tag{B.9}$$

It follows from Bonferroni's inequality, inequalities (B.8) and (B.9) that,

$$\begin{aligned}
&\Pr\left(\left|\widehat{S}_{k2}-S_{k2}\right|\geq 3\varepsilon\right)=\Pr\left(\left|\widehat{S}_{k2,1}\widehat{S}_{k2,2}-S_{k2,1}S_{k2,2}\right|\geq 3\varepsilon\right)\\
\leq&\Pr\left\{\left|(\widehat{S}_{k2,1}-S_{k2,1})S_{k2,2}\right|\geq\varepsilon\right\}+\Pr\left\{\left|S_{k2,1}(\widehat{S}_{k2,2}-S_{k2,2})\right|\geq\varepsilon\right\}\\
&+\Pr\left\{\left|(\widehat{S}_{k2,1}-S_{k2,1})(\widehat{S}_{k2,2}-S_{k2,2})\right|\geq\varepsilon\right\}\\
\leq&8\exp\left\{-\varepsilon^2 n^{1-2\gamma}/(16C^2)\right\}+8nC\exp\left(-sn^{2\gamma}/4\right),
\end{aligned}\tag{B.10}$$

where the last inequality holds when ε is sufficiently small and C is sufficiently large.

It remains to the uniform consistency of $\widehat{S}_{k3}$. We first study the following U-statistic:

$$\begin{aligned}
\widehat{S}^*_{k3} &= \frac{1}{n(n-1)(n-2)}\sum_{i<j<l}\Big\{\|X_{ik}-X_{jk}\|_1\|\mathbf{y}_j-\mathbf{y}_l\|_q+\|X_{ik}-X_{lk}\|_1\|\mathbf{y}_j-\mathbf{y}_l\|_q+\\
&\qquad\|X_{ik}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_l\|_q+\|X_{lk}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_l\|_q+\\
&\qquad\|X_{lk}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q+\|X_{lk}-X_{ik}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q\Big\}\\
&=: \frac{6}{n(n-1)(n-2)}\sum_{i<j<l}h_3(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j;X_{lk},\mathbf{y}_l).
\end{aligned}\tag{B.11}$$

Here, $h_3(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j;X_{lk},\mathbf{y}_l)$ is the kernel of U-statistic $\widehat{S}^*_{k3}$. Following the arguments to deal with $\widehat{S}^*_{k1}$, we decompose h_3 into two parts: $h_3=h_3\mathbf{1}(h_3>M)+h_3\mathbf{1}(h_3\leq M)$. Accordingly,

$$\begin{aligned}
\widehat{S}^*_{k3} &= \frac{6}{n(n-1)(n-2)}\sum_{i<j<l}h_3\mathbf{1}(h_3\leq M)+\frac{6}{n(n-1)(n-2)}\sum_{i<j<l}h_3\mathbf{1}(h_3>M)\\
&= \widehat{S}^*_{k3,1}+\widehat{S}^*_{k3,2},\\
S_{k3} &= E\left\{h_3\mathbf{1}(h_3\leq M)\right\}+E\left\{h_3\mathbf{1}(h_3>M)\right\}=S_{k3,1}+S_{k3,2}.
\end{aligned}$$

Following similar arguments for proving (B.2), we can show that

$$\Pr\big(\left|\widehat{S}^*_{k3,1}-S_{k3,1}\right|\geq\varepsilon\big)\leq 2\exp\left(-2\varepsilon^2 m'/M^2\right),\tag{B.12}$$

where $m' = [n/3]$ because $\widehat{S}^*_{k3,1}$ is a third-order U-statistic.

Then we deal with $\widehat{S}^*_{k3,2}$. We observe that $h_3(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j; X_{lk}, \mathbf{y}_l) \leq 4(X^2_{ik} + X^2_{jk} + X^2_{lk} + \|\mathbf{y}_i\|^2_q + \|\mathbf{y}_j\|^2_q + \|\mathbf{y}_l\|^2_q)/6$, which will be smaller than M if $X^2_{ik} + \|\mathbf{y}_i\|^2_q \leq M/2$ for all $1 \leq i \leq p$. Thus, for any $\varepsilon > 0$, the events satisfy

$$\left\{\left|\widehat{S}^*_{k3,2}\right| > \varepsilon/2\right\} \subseteq \left\{X^2_{ik} + \|\mathbf{y}_i\|^2_q > M/2, \text{ for some } 1 \leq i \leq p\right\}.$$

By using the similar arguments to prove (B.5), it follows that

$$\Pr\left(\left|\widehat{S}^*_{k3,2} - S_{k3,2}\right| > \varepsilon\right) \leq \Pr\left(\left|\widehat{S}^*_{k3,2}\right| > \varepsilon/2\right) \leq 2nC\exp(-sM/4). \tag{B.13}$$

Then, we combine the results (B.12) and (B.13) with $M = cn^\gamma$ for some $0 < \gamma < 1/2 - \kappa$ to obtain that

$$\Pr\left(\left|\widehat{S}^*_{k3} - S_{k3}\right| \geq 2\varepsilon\right) \;\leq\; 2\exp\left(-2\varepsilon^2 n^{1-2\gamma}/3\right) + 2nC\exp\left(-sn^\gamma/4\right). \tag{B.14}$$

By the definition of $\widehat{S}_{k3}$,

$$\widehat{S}_{k3} = \frac{(n-1)(n-2)}{n^2}\left\{\widehat{S}^*_{k3} + \frac{1}{(n-2)}\widehat{S}^*_{k1}\right\}.$$

Thus, using similar techniques to deal with $\widehat{S}_{k1}$, we can obtain that

$$\begin{aligned}\Pr\left(\left|\widehat{S}_{k3} - S_{k3}\right| \geq 4\varepsilon\right) \;=\; \Pr\Bigg\{\Bigg|&\frac{(n-1)(n-2)}{n^2}\left(\widehat{S}^*_{k3} - S_{k3}\right) - \frac{3n-2}{n^2}S_{k3} \\ &+ \frac{n-1}{n^2}\left(\widehat{S}^*_{k1} - S_{k1}\right) + \frac{n-1}{n^2}S_{k1}\Bigg| \geq 4\varepsilon\Bigg\}.\end{aligned}$$

Using similar arguments for dealing with S_{k1}, we can show that S_{k3} is uniformly bounded in

p. Taking n large enough such that $\{(3n-2)/n^2\}S_{k3} \leq \varepsilon$ and $\{(n-1)/n^2\}S_{k1} \leq \varepsilon$, then

$$\begin{aligned}\Pr\big(\big|\widehat{S}_{k3} - S_{k3}\big| \geq 4\varepsilon\big) &\leq \Pr\big(\big|\widehat{S}^*_{k3} - S_{k3}\big| \geq \varepsilon\big) + \Pr\big\{\big|\widehat{S}^*_{k1} - S_{k1}\big| \geq \varepsilon\big\} \\ &\leq 4\exp\left(-\varepsilon^2 n^{1-2\gamma}/6\right) + 4nC\exp\left(-sn^{\gamma}/4\right).\end{aligned} \tag{B.15}$$

The last inequality follows from (B.6) and (B.14). This, together with (B.6), (B.10) and the Bonferroni's inequality, implies

$$\begin{aligned}&\Pr\big\{\big|(\widehat{S}_{k1} + \widehat{S}_{k2} - 2\widehat{S}_{k3}) - (S_{k1} + S_{k2} - 2S_{k3})\,\big| \geq \varepsilon\big\} \\ \leq&\Pr\big(\big|\widehat{S}_{k1} - S_{k1}\big| \geq \varepsilon/4\big) + \Pr\big(\big|\widehat{S}_{k2} - S_{k2}\big| \geq \varepsilon/4\big) + \Pr\big(\big|\widehat{S}_{k3} - S_{k3}\big| \geq \varepsilon/4\big) \\ =&O\left\{\exp\left(-c_1\varepsilon^2 n^{1-2\gamma}\right) + n\exp\left(-c_2 n^{\gamma}\right)\right\},\end{aligned} \tag{B.16}$$

for some positive constants c_1 and c_2. The convergence rate of the numerator of $\widehat{\omega}_k$ is now achieved. Following similar arguments, we can obtain the convergence rate of the denominator. In effect the convergence rate of $\widehat{\omega}_k$ has the same form of (B.16). We omit the details here. Let $\varepsilon = cn^{-\kappa}$, where κ satisfies $0 < \kappa + \gamma < 1/2$. We thus have

$$\begin{aligned}\Pr\Big\{\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\Big\} &\leq p\max_{1\leq k\leq p}\Pr\left\{|\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\right\} \\ &\leq O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right).\end{aligned}$$

The first part of Theorem 1 is proven.

Now we deal with the second part of Theorem 1. If $\mathcal{D} \nsubseteq \widehat{\mathcal{D}}^{\star}$, then there must exist some $k \in \mathcal{D}$ such that $\widehat{\omega}_k < cn^{-\kappa}$. It follows from condition (C2) that $|\widehat{\omega}_k - \omega_k| > cn^{-\kappa}$ for some $k \in \mathcal{D}$, indicating that the events satisfy $\big\{\mathcal{D} \nsubseteq \widehat{\mathcal{D}}^{\star}\big\} \subseteq \big\{|\widehat{\omega}_k - \omega_k| > cn^{-\kappa},\ \text{for some } k \in \mathcal{D}\big\}$,

and hence $\mathcal{E}_n = \{\max_{k\in\mathcal{D}} |\widehat{\omega}_k - \omega_k| \leq cn^{-\kappa}\} \subseteq \{\mathcal{D} \subseteq \widehat{\mathcal{D}}^\star\}$. Consequently,

$$\begin{aligned}
\Pr(\mathcal{D} \subseteq \widehat{\mathcal{D}}^\star) &\geq \Pr(\mathcal{E}_n) = 1 - \Pr(\mathcal{E}_n^c) = 1 - \Pr\big(\min_{k\in\mathcal{D}} |\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\big) \\
&= 1 - s_n \Pr\left\{|\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\right\} \\
&\geq 1 - O\left(s_n\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right),
\end{aligned}$$

where s_n is the cardinality of $\mathcal{D}$. This completes the proof of the second part. □

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., et al. (2000), "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," *Nature Genetics*, **25**, 25-29.

Bild, A., Yao, G., Chang, J. T., Wang, Q., Potti, A., et al. (2006), "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature* **439** 353-357.

Candes, E. and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n (with discussion)," *Annals of Statistics*, **35**, 2313–2404.

Chen, L. S., Paul, D., Prentice, R. L. and Wang, P. (2011), "A regularized Hotelling's T^2 test for pathway analysis in proteomic studies," *Journal of the American Statistical Association* **106** 1345–1360.

Efron, B., Hastie T., Johnstone, I. and Tibshirani, R. (2004), "Least angle regression (with discussion)," *Annals of Statistics*, **32**, 409–499.

Efron, B., and Tibshirani, R. (2007), "On Testing the Significance of Sets of Genes," *The Annals of Applied Statistics*, **1**, 107-129.

Fan, J., Feng, Y. and Song, R. (2011), "Nonparametric independence screening in sparse ultra-high dimensional additive models," *Journal of the American Statistical Association*, **106**, 544–557.

Fan, J., and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and it oracle properties," *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space (with discussion)," *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009), "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of Machine Learning Research*, **10**, 1829–1853.

Fan, J. and Song, R. (2010), "Sure independence screening in generalized linear models with NP-dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

Hall, P. and Miller, H. (2009), "Using generalized correlation to effect variable selection in very high dimensional problems," *Journal of Computational and Graphical Statistics*, **18**, 533–550.

Ji, P. and Jin, J. (2012), "UPS delivers optimal phase diagram in high dimensional variable selection," *Annals of Statistics*, **40**, 73-103.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., et al. (2008), "Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses," *Science*, **321** 1801.

Kim, Y., Choi, H. and Oh, H. S. (2008), "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association,* **103**, 1665–1673.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., et al. (2003), "PGC-1-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes," *Nature Genetics*, **34**, 267-273.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression approach for microarray data analysis," *Journal of Computational Biology,* **10**, 961–980.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics,* New York: John Wiley & Sons Inc.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., et al. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences of the USA*, **102**, 15545-15505.

Székely, G. J. and Rizzo, M. L. (2009), "Brownian distance covariance," *Annals of Applied Statistics*, **3**, 1233–1303.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, **35**, 2769–2794.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005), "Discovering Statistically Significant Pathways in Expression Profiling Studies," *Proceedings of the National Academy of Sciences of the USA*, **102**, 13544-13549.

Tibshirani, R. (1996), "Regression shrinkage and selection via LASSO," *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Wang, H. (2009), " Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, **104**, 1512–1524.

Zhao, S. D. and Li, Y. (2012), "Principled sure independence screening for Cox models with ultra-high-dimensional covariates," *Journal of Multivariate Analysis*, **105**, 397–411.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011), "Model-free feature screening for ultrahigh dimensional data," *Journal of the American Statistical Association*, **106**, 1464–1475.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, **36**, 1509–1533.

Zou, H. and Zhang, H. H. (2009), "On the adaptive elastic-net with a diverging number of parameters," *Annals of Statistics*, **37**, 1733-1751.